Reviewer 2

The submitted paper discusses an interesting, novel and computationally efficient approach to model evaluation which builds on concepts of ensemble Kalman filtering and analog approaches to build skill scores which indicate some degree of skill in identifying model errors when assessed in a perfect model framework. The technique is demonstrated for a highly idealised case (the Lorenz 63 model) and an intermediate complexity climate model.

The paper is well written, and novel. My opinion is that it should be published with only minor edits.

We would like to thank the reviewer for the comments that helped to improve the clarity of the manuscript.

Minor comments:

1 The approach described here is acceptable as a proof of concept - however it is likely not an optimal use of the data used in the training simulation used to assemble the analogs. In particular, the use of only small-scale information in the construction of analogs is discarding valuable information which would be represented in the covariance structure of the model output. The need to minimise the state space of model in order to find acceptable analogs is clear - but my suspicion is that a compression of state space which preserves elements of large scale covariance (such as PCA), rather than isolated regional analyses, would be even more effective.

We would like to thank the reviewer for bringing this interesting point. We agree with this idea and we add the following sentence in the discussion section:

“Implementing the combination of CME and AnDA in real-data cases brings additional challenges. For instance, in this work the application of the analog regression technique to a high-dimensional problem is achieved by using local domains. However, this approach does not take advantage of the covariance structure of the model output. This structure could be retrieved through a principal component analysis which may allow the implementation of the analog regression in a low dimensional space while keeping the main aspects of large scale circulation patterns.”

2. A discussion of the dependency of performance on training run length for the analog would be useful, compared to a forecast-based approach (in the Lorenz case), and in terms of the ability to distinguish model errors (for SPEEDY).

A sensitivity experiment to the length of the catalog has been performed with the Lorenz 63 system, with or without data assimilation (see Table 1 for details). We also add a discussion about the selection of the length of the catalog in the SPEEDY experiments:

“AnDA experiments are conducted assimilating the observations generated from the last three years of the TRUE simulation. The catalogs for the analog forecasting are constructed from the first 25 years of the RH08, RH07, and TRUE model runs and 250~analogs are used for the forecast.

In the SPEEDY experiments, the catalog contains over 36.000 samples (which is almost 4 times the size of the largest catalog which we tried with the Lorenz model). Although the local state space dimension that we used in SPEEDY is much larger (27 grid points), we argue that since there are substantial correlations among the state variables, the effective dimension can be significantly smaller.

The number of ensemble members is 30. To increase the evidence associated with the local dynamics of the models the assimilation frequency is set to 24 hours. To take advantage of 6-hourly data, at each local domain, we perform four DA experiments which are run independently from each other starting at 00, 06, 12 and 18 UTC on the first day. These four DA cycles are performed over the same 3-years period. These configuration settings have been chosen based on preliminary experiments performed over a limited number of local domains in which the sensitivity of the results to these parameters has been explored.

The analysis obtained from these experiments are merged to obtain a total of 4,380 analysis cycles over the three-years assimilation period (4 DA experiments x 1095 cycles each). It is important to note that the generation of the catalog brings a significant computational cost in this approach since it requires running the global numerical model over a long period of time. However, we argue that for the implementation of this technique in real data applications, available long model simulations like those produced by the Coupled Model Intercomparison Project (Eyring et al. 2016) can be used. Moreover, the length of these catalogs are of the same order of magnitudes as the ones used in the idealized experiments with the SPEEDY model.”

4. Though the authors have demonstrated that CME provides a generally improved regional assessment of model error, this is not universally the case - especially for RH08, where ME provides a stronger signal in a number of regions. A short discussion on regions where this occurs, and potentially why, would be useful.

We agree with the reviewer, Figure 8 shows some areas where in fact RMSE performs better than CME at selecting the perfect model. We add a comment on this on the result section as well as a brief discussion of the possible cause of this issue.

“Although CME usually performs better than the RMSE at identifying the correct model, this is not always the case (see for example in Figure 8 how the probability of correct identification is larger for the RMSE than for CME near the Equator). This result may be due to an overestimation of the forecast error covariance $\boldsymbol{\Sigma}^f$, computed within the analog procedure. Indeed, as explained in Eq. (11), an augmentation of this error matrix implies a diminution of the CME, and thus a decrease of performance of this metric.”